Identifying the most reliable and valid bladder health screening tool: a systematic review

Booth, Lorna; Skelton, Dawn A.; Hagen, Suzanne; Booth, Jo

Published in:
Disability and Rehabilitation

DOI:
10.1080/09638288.2018.1561953

Publication date:
2020

Document Version
Author accepted manuscript

Identifying the Most Reliable and Valid Bladder Health Screening Tool: a Systematic Review

Bladder Health Screening Tools

Lorna Booth[1]; Dawn A Skelton[1]; Suzanne Hagen[2]; Joanne Booth[1].

[1] *School of Health and Life Sciences, Glasgow Caledonian University, Cowcaddens Road, Glasgow, G4 0BA.*

[2] *Nursing, Midwifery and Allied Health Professions Research Unit, Glasgow Caledonian University, Cowcaddens Road, Glasgow, G4 0BA.*

Corresponding Author: Lorna Booth, School of Health and Life Sciences
Glasgow Caledonian University, Cowcaddens Road, Glasgow, UK, G4 0BA.
Tel: +44(0) 141 2731358
Email: lorna.booth@gcu.ac.uk
ORCid: 0000-0001-5776-3541

Abstract:

Lower Urinary Tract Symptoms are common in advancing age and a major cause of disability through avoidance of activity and social engagement. This systematic review aimed to identify the most valid and reliable brief screening tool for these symptoms or bladder problems, to incorporate into a health promotion programme for older adults to facilitate discussion about self-management.

Review eligibility criteria included studies published between 1990 and November 2018, reporting the validity, reliability and/or acceptability of bladder health screening tools. Six electronic databases were searched.

Twenty-two studies were included. Three screening tools met the criteria: International Prostate Symptom Score; International Consultation on Incontinence Questionnaire Urinary Incontinence Short-Form; Bladder Control Self-Assessment Questionnaire.

Test-retest reliability for total scores of the International Prostate Symptom Score and International Consultation on Incontinence Questionnaire Urinary Incontinence Short-Form was acceptable. All three questionnaires showed evidence of acceptable levels of internal consistency and of convergent validity.

Having favourable psychometric scores compared to the Bladder Control Self-Assessment Questionnaire and for ease of use and trustworthiness of a simple questionnaire, the International Prostate Symptom Score and International Consultation on Incontinence Questionnaire Urinary Incontinence Short-Form met the criteria for recommendation for raising awareness and bladder health promoting interventions to reduce associated disability.

Keywords: bladder health, urinary incontinence, questionnaire, validity, reliability.

Introduction

Lower Urinary Tract Symptoms (LUTS) are indicators of poor bladder health. LUTS can be split into three categories; storage, voiding and post micturition symptoms [1]. Storage symptoms include frequency, nocturia, urgency, urge urinary incontinence, stress urinary incontinence, mixed urinary incontinence, and other urinary incontinence; voiding symptoms include intermittency, slow stream, straining, and terminal dribble; and post micturition symptoms present as incomplete emptying and post micturition dribble [2]. Therefore, an individual may experience LUTS alone whilst remaining continent or urinary symptoms may be accompanied by urinary incontinence, defined by the International Continence Society as "the complaint of involuntary loss of urine" [3].

Bladder health deterioration, including neurogenic bladder dysfunction caused by neurological conditions, can have detrimental effects on the individual's physical and psychological state [4-7]. Urinary incontinence is a disabling condition and although not life-threatening, in older adults it can lead to skin breakdown, frailty, social exclusion, psychological stress, poorer quality of life, financial burden and poorer physical mobility [8-11]. Urinary incontinence in older adults also increases the risk of falls and fractures [12,13] and is associated with more hospital admissions [14]. Usual daily living activities can be difficult or impossible as a consequence of the inability to maintain continence [15]. It has been reported that urinary incontinence is a major cause of independence loss [16], and is one of the most frequent causes for care home admissions [17] with prevalence rates as high as 77% in nursing homes residents [18]. Many of these consequences of urinary incontinence have been defined as the global outcomes of disability [19] and often occur through avoidance of activity and social engagement as a result of urinary incontinence [20-23].

There is a considerable prevalence of LUTS and urinary incontinence among the general population. Over 66 percent of women and 62 percent of men aged forty years or

more reported having at least one LUTS in a large scale multinational population-based survey study [24]. Global prevalence of urinary incontinence was estimated at 348 million people in 2008 and this number is expected to rise year on year [2]. Urinary incontinence incidence estimates for the UK, range from 1.6 to 69**%** in females and 2.2 to 25% in males [25], varying due to variations in measurement and definitions [26] and differences in populations [27].

Unfortunately, urinary incontinence is often seen as part of "normal ageing", with the view that nothing can be done about it, and as an embarrassing issue associated with negative stigma [18-32]. Only fifty percent of older adults experiencing incontinence are expected to seek help from healthcare practitioners [33], yet most mild urinary incontinence can be easily managed or cured using simple lifestyle and behavioural interventions [34,35]. Promoting understanding of bladder health and supporting self-management as part of a health promotion intervention could provide an opportunity not only to detect early bladder health issues, but also help prevent the progression of LUTS to urinary incontinence [36,37].

Parsons et al. [38] examined the progression of LUTS in community-dwelling older men, and found that 29% who had initially presented with no or mild LUTS reported clinically significant LUTS two years later. More recently, research has shown that more than 50% of middle-aged and older men experience worsening of LUTS over a 3-year period [39]. This highlights the importance of early screening of LUTS to recognise potential problems, to enable understanding and provide support for self-management of symptoms.

Previous reviews have been conducted to establish the quality of bladder symptom severity assessments and/or associated quality of life [40]. However, there has been no such review considering the use of such tools for a general, non-clinical population of men and women, for the purposes of early detection and raising awareness about bladder health. Hence, there is a need to identify a simple, reliable, valid and acceptable method to screen for

LUTS and urinary incontinence which can be added to health promotion interventions to promote bladder health.

The quality of a structured questionnaire and its applicability to practice are important considerations when selecting a potential screening tool to use in research. A questionnaire's quality should be assessed by its measurement properties, usually through two specific psychometric concepts of reliability and validity [41]. It has also been recommended that health questionnaires should be brief and easy to use to be applicable to practice and to enhance response rates [42,43].

The aim of the current systematic review is to identify the most psychometrically robust tool, for bladder health screening which is applicable to both men and women, for inclusion in a health promotion intervention suitable for a non-clinical population.

Methods

The review was developed to answer the following research questions:

(1) Which generic bladder health screening questionnaires have been evaluated for their psychometric properties for use within a male and female adult population?

(2) Which of the identified screening questionnaires are brief and easy to use (assessed as having ten items or less) and have a low level of missing data?

(3) Which of the included screening questionnaires (as per question 1 and 2), have the best psychometric properties for use with a non-clinical adult population?

This systematic review was built on a robust protocol, which adhered to the Preferred Reporting Items for Systematic Reviews and Meta-Analyses Protocols (PRISMA-P) guidelines [44].

Information Sources

Relevant papers were identified through systematic searching of the following electronic databases: Medline (EBSCO SP), CINAHL (EBSCO SP), AMED (EBSCO SP), PsycInfo (ProQuest SP), EMBASE (OVID SP) and Web of Science (Thomson Reuters SP). Reference lists of included studies were also scanned to allow for literature saturation. A manual search was conducted in a key journal in the field (Neurourology and Urodynamics), searching specifically for commonly-known bladder health screening questionnaires, as advised by a urinary health expert, and other tools identified through the database search. PROSPERO, DARE and Google Scholar were searched for recently-completed reviews, to ensure no reviews similar to the current one had been published.

The systematic search adopted a strategy which combined selected subject headings and keywords consisting of three parts: urinary bladder health (e.g. LUTS; urinary incontinence), screening tools (e.g. questionnaire) and psychometrics (e.g. reliability; validity). A limiter of 'adults' was applied to each search. Before the strategies were finalised, the search keywords were checked against each individual database's medical subject headings (MESH) or thesaurus terms. The search strategy for MEDLINE is shown in Supplementary table 1. Some of these words/terms altered slightly, depending on the database being searched.

Date limits of 1990 to November 2018 and English language limiters were applied. Articles assessing validity or reliability of translated versions of English language bladder health screening questionnaires were included.

Inclusion Criteria

- Peer reviewed primary quantitative research studies involving adults aged 18 years and over.
- Studies that specifically report on reliability and/or validity of generic bladder health (LUTS and/or urinary incontinence) questionnaires.

- Full-text published studies written in English.
- Studies undertaken in any type of setting/context.

Exclusion Criteria

- Studies of questionnaires which are specific to a particular sub-type of urinary incontinence or symptom constellation.
- Studies of questionnaires only applicable to either men or women.
- Studies of questionnaires measuring Quality of Life and not bladder symptoms.

Study Selection Strategy

Two review authors (LB, JB) independently screened the title and abstract data using the pre-developed test screening questions based on eligibility, inclusion and exclusion criteria. Any disagreement was discussed and agreement reached, with the option of a third reviewer if needed. Full text reports were obtained for titles and abstracts that met the eligibility criteria or if there was any uncertainty. Reasons for exclusion are documented in figure 1.

Data Extraction

A data extraction form (Supplementary table 2) was used to collect details of the screening questionnaire identified, including evidence of reliability and validity for each, and applicability to practice measures. As definitions of different psychometric properties are often used interchangeably among different authors, each definition was operationalised for the current review (table 1). Data were extracted independently from all eligible studies by one reviewer (LB) and validated by a second reviewer (JB). Disagreements were resolved by discussion.

[insert table 1 here]

Quality Assessment

In the absence of a standardised quality appraisal tool applicable to screening tools, an adapted appraisal checklist (from Bellet et al. [45]; Supplementary table 3) was used to assess

the overall methodological quality of the included studies. Bellet et al. [45] used the original checklist to assess the quality of articles exploring validity, reliability and responsiveness of objective clinical tools. The adapted checklist did not include items related to responsiveness. A comparison between Bellet et al's [45] appraisal items and the adapted appraisal items, can be found in Supplementary table 3. The methodological quality of identified papers was assessed by one reviewer (LB) and validated by a second reviewer (JB). Disagreements were resolved by discussion.

Bellet et al. [45] suggested a 60% positive response to the relevant checklist items indicated high quality whereas less than 40% response indicated poor quality. Several items were not applicable to all of the included studies, particularly if a standard reference had not been used for comparison within the study. Therefore, it was decided that these cut-off points would be very arbitrary and lack meaning in this review as each item would have a different weighting dependant on how many items were relevant to the specific study, thus providing an unfair comparison.

Methods of Analysis/Synthesis: Acceptable levels

Acceptable levels of reliability and validity scores were defined using the following criteria:

Reliability Measures:

- Intra-class correlation coefficient ≥ 0.70 (ICC) [46-49].
- Kappa ≥ 0.70 [48,49].
- Cronbach's alpha ≥ 0.70 [46].
- Pearson's correlation or Spearman's rank ≥ 0.80 [46,48,49].

Validity Measures:

- Correlation coefficients ≥ 0.50 [46].
- Sensitivity ≥ 80%; Specificity ≥ 60% [46].

- Receiver Operating Characteristic (ROC) curves: Area Under the Curve (AUC) ≥ 0.80 [46].

The practicability and acceptability of these tools were also assessed for the purpose of using the tool as part of a bladder health promotion programme by considering the number of items included in the questionnaire, time taken to complete the questionnaire, the level of missing data reported and the dropout/response rate reported.

Results

Results from Screening and Selection Process:

Twenty-three articles were included in the review. The results of the selection and screening process are presented in the PRISMA flowchart (figure 1). Twenty-three articles met the inclusion criterion, however two of the included articles were identified as companion studies and were treated as one study throughout the review [50,51].

[insert figure 1 here]

Applicability to Practice

Three bladder health questionnaires that fulfilled the eligibility criteria were identified: The International Prostate Symptom Score (IPSS)/The American Urological Association Symptom Index (AUASI); The International Consultation on Incontinence Questionnaire-Urinary Incontinence Short Form (ICIQ-UI SF); and The Bladder Control Self-assessment Questionnaire (B-SAQ). The three questionnaires were assessed for applicability to practice within a community health promotion programme (table 2).

Lower urinary tract symptoms but not urinary leakage, are screened for by the IPSS/AUASI, whereas urinary incontinence only is screened for by the ICIQ-UI SF. The B-SAQ screens for LUTS and urinary incontinence. The IPSS and the B-SAQ have a total of eight items and the ICIQ-UI SF has four items, thus they are short to complete. The IPSS/AUASI has seven items that measure urinary symptom frequency and severity and

provides a total score indicating overall severity of LUTS experienced. The B-SAQ has four items that measure symptoms and urinary incontinence and four items that are related 'bother' questions, all of which are included in the total score. The ICIQ-UI SF has three questions included in the total score, measuring frequency, amount and impact of leakage, with an additional unscored self-diagnosis question for type of urinary incontinence experienced. Scoring systems for the three identified questionnaires involve simple summation. The time taken to complete the questionnaire was only reported for the B-SAQ and was less than five minutes.

All three questionnaires were reported to be easy to understand and results from several studies show a high percentage of participants being able to complete them correctly with low levels of missing data (table 2) [51,58,64,66-70,73,74]. However, Cam et al. [52] reported the IPSS to be complicated as they found that more than half of patients were unable to correctly complete the questionnaire in full. They identified that lower levels of education greatly affected how easily the questionnaire is understood. Those who reported lower education levels were three times more likely to be unable to complete. This was supported by other studies that found high numbers of people who found the IPSS difficult to understand [52-54]. However, other studies have shown the IPSS to be easy to understand [50].

[insert table 2 here]

Study Characteristics

Of the three identified screening questionnaires, the most frequently psychometrically assessed questionnaire was the IPSS, as demonstrated in table 3 [n = 11; 3 x AUASI [50,55,56]; 8 x IPSS [52,57-63], where five articles assessed the English version of the tool and six articles assessed other language versions including: Mandarin [59], Turkish [52], Spanish [61,62], Japanese [58] and Arabic [57]. Nine articles assessed the psychometric

properties of the ICIQ-UI SF [64-72], where five assessed the English version [64,69-71] and five articles explored alternative language versions including: Arabic [67], Slovene [66], Japanese [65], Italian [68], Chinese and Malay [64]. Only two articles were found which assessed the validity and reliability of the B-SAQ [73,74], both of which used English versions of the tool.

The total sample inclusive of all the studies is 7901 participants (table 3), with the highest number of people being used to assess the IPSS (n = 4180) followed by the ICIQ-UI SF (n = 3221) and the B-SAQ (n = 540). Sample sizes of individual studies ranged from 57 participants to 1620 participants with mean ages ranging from 38 years to 86 years. The majority of studies used a mean sample of people in their sixties.

Most studies (n = 11) included only men. This is not surprising as most of these studies (n = 9) assessed the IPSS, which was developed to screen for Benign Prostatic Hyperplasia (BPH). Seven studies assessed data on both men and women [58,65-67,69,71,72]. Of these five studies, one assessed the IPSS [58] and the remaining six studies assessed the ICIQ-UI SF. Five of the six studies had a higher female to male ratio (60%-40%; 84%-16%; 61%-39%; 66%-34%; 77%-23%) respectively [65-67,69,71], (the remaining study did not report the gender ratio [80]), whereas the only study assessing the IPSS for both sexes had a higher male to female ratio (54%-46%). Three studies reported data on women only [64,68,71]. One study did not specifically state what sex their participants were [52]. Although the study by Cam et al. [52] assessed the IPSS it would be inappropriate to assume that the sample included only males as the health status of their inclusion criteria was patients with lower urinary tract symptoms and not specifically BPH, therefore the sample in the study was considered to be men and women. No attempt was made to contact the authors of the Cam et al. [52] paper due to the prolonged interval since publication (more than 10 years).

The bladder condition reported in most cases was LUTS and/or urinary incontinence. Controls were often people who had no history of LUTS and/or urinary incontinence.

Seventeen of the 22 studies were conducted in patients attending urology clinics or other specialist clinics. One study included a specialist sample receiving secondary care and a non-specialist sample receiving primary care (ICIQ-UI SF) [66], another study included only a non-specialist sample receiving primary care (IPSS) [58]. Gotoh et al. [65] did not state where their sample was recruited, or where the screening took place. The ICIQ-UI SF and the IPSS have both been psychometrically assessed in specialist and non-specialist health contexts (table 3). The B-SAQ has been assessed only in specialist contexts [73,74].

[insert table 3 here]

Reliability and Validity Results

All but two articles [71, 73] assessed the reliability of the tool (table 4); test-retest reliability (n = 16), internal consistency (n = 14), inter-rater reliability (n = 4). Twelve studies explored the tool's validity (table 5); content validity (n = 5), criterion validity (n = 3), construct validity (n = 11). Three of the studies which explored construct validity, actually reported their results as criterion validity [60,65,74]. As no objective criterion was used for comparison or agreement, it was decided that the results of these three studies should be reported as construct validity in accordance with the stated validity definitions (table 1). Gotoh et al. [65] claim to have assessed discriminant validity in their study, however they used an objective criterion (1hr pad test) hence this assessment was operationalised as criterion validity. Tables 4 and 5 include summaries of the reliability and validity results respectively, for the studies included in the current review.

Test-retest reliability

Of the sixteen studies that assessed test-retest reliability, seven assessed the ICIQ-UI SF, eight assessed the IPSS or AUASI, and one assessed the B-SAQ (table 4).

One study assessing test-retest reliability for the IPSS, using Pearson's correlation, did not provide the statistical result [63], hence was excluded. Of the remaining seven studies two studies used Pearson's correlations, reporting values of 0.81 and 0.92 respectively [56,60]. Four studies used the Intra-class correlation coefficient (ICC) to assess test-retest reliability of the IPSS total 8-item score (including 7 symptom items and 1 quality of life item), reporting values ranging from 0.76 to 0.93 [51,57,59,61]. One study reported the ICC for the 7 symptom items only which yielded a result of 0.87 [62]. Therefore, test-retest reliability for the total scores for the IPSS is considered acceptable as the results consistently evidence an ICC score ≥ 0.7 and Pearson's correlation ≥ 0.80 [46-49].

Of the seven studies that assessed test-retest reliability for the ICIQ-UI SF (table 4), four studies used Kappa coefficients, two for total score [67,69] and two for each of the three individual items [66,72], two used ICCs [64,68], and one used Kappa coefficient for item 1 and 2 and ICC for item 3 and total [65]. Kappa results for total scores ranged from 0.74 to 0.85. Kappa results for item 1 and 2 ranged from 0.61 to 0.99 and 0.62 to 0.98 respectively. ICC results for total scores ranged from 0.91 to 0.96. Therefore, test-retest reliability for the total scores for the ICIQ-UI SF is considered to be acceptable as the results evidence Kappa and ICC total scores ≥ 0.7 [46-49]. The test-retest reliability of individual items however is less consistent. Notably, the lower Kappa coefficient results for items 1 and 2 were reported in a study testing the validity and reliability of the Japanese version of ICIQ-SF [65]. Item 3 was also found to have a lower than acceptable Kappa result on one study [73].

Only one study [74] has assessed the test-retest reliability of the B-SAQ total scores which yielded a Kappa coefficient result of 0.60 to 0.69 (table 4), indicating an unacceptable level and poorer test-retest reliability than the IPSS and the ICIQ-SF. Basra et al. [74] also reported Lin's coefficient of concordance ranging from 0.86 and 0.99 which indicates high

levels of test-retest reliability, however this measure has not been used in the assessment of any other tools or included studies, hence a lack of available comparability.

The most common timescale between test and retest was 1 week (n = 7), followed by 2 weeks (n = 4) and over 2 weeks (n = 2). Assessments of the test-retest for the ICIQ-UI SF used a 2-week timescale more often (n = 3) whereas the assessments for the IPSS used a 1-week assessment more often (n = 5). The one test-retest assessment for the B-SAQ used a 4-week timescale.

Internal Consistency

Of the fourteen studies that assessed internal consistency, six assessed the IPSS/AUASI, seven assessed the ICIQ-UI SF and one assessed the B-SAQ (table 4). The reported Cronbach's Alphas for the full IPSS, ICIQ-UI SF and B-SAQ ranged from 0.75 to 0.97, 0.60 to 0.92, and 0.91 respectively, indicating that the IPSS and the B-SAQ may have more desirable levels of internal consistency than the ICIQ-UI SF. Notably, the lower levels of internal consistency for the ICIQ-UI SF were found in only two of the seven articles [64,70]. Excluding these findings, all three questionnaires have evidence of acceptable levels of internal consistency (≥ 0.7).

Inter-rater Reliability

Four studies assessed inter-rater reliability of the IPSS using; Pearson Correlation [63], ICC [56], Kappa and Spearman Correlation [55], and paired Wilcoxon tests [56], which either found no significant differences between physician/interviewer administered scores and self-administered scores or significant agreement between scores (table 4). None of the included studies assessed inter-rater reliability for the ICIQ-UI SF or the B-SAQ. Therefore, the IPSS has the most evidence relating to inter-rater reliability compared to the ICIQ-UI SF and the B-SAQ, and that evidence suggests inter-rater reliability for the IPSS is good.

[insert table 4 here]

Content Validity

Both the ICIQ-UI SF and the B-SAQ have been assessed for face and logical content validity (table 5), with the conclusion that all items of both questionnaires were relevant. Some studies report content validity for the IPSS by measuring the levels of missing data however, this review takes the same stance as previous research [75] that missing data is a measure of acceptability rather than face or logical content validity.

Criterion Validity

The ICIQ-UI SF is the only questionnaire of the three identified in this review to have been assessed for criterion validity using an objective "gold standard" measurement (table 5). One study compared responses to question 6 of the ICIQ-UI SF (perceived cause of leakage) with urodynamic testing diagnoses which yielded an acceptable Kappa coefficient correlation of 0.77 [66]. Another study compared the ICIQ-UI SF with a 1-hour pad test and found that frequency of leakage, amount of leakage and total ICIQ-UI SF scores were significantly correlated with the 1-hour pad test results [65]. Mary-Heck et al [71] found that the ICIQ-UI-SF had 100 percent specificity and sensitivity when compared diagnostic results made by a team of health professionals. Tubaro et al. [68] however found that there was a large variability between the scores on the ICIQ-UI SF and the reported results from a 72-hour voiding diary, although specific results were not provided.

Construct Validity: Discriminative Validity

The ICIQ-UI SF has been found to differentiate (discriminative validity) between disparate groups of people including: males and females for prevalence of UI [69]; type of urinary incontinence [66,67], for perceived cause of urinary incontinence [69], and also between cases and controls for prevalence of urinary incontinence [68,69] and specifically for stress urinary incontinence [64]. The ICIQ-UI SF has also been shown to discriminate between different types of urinary incontinence in regards to the associated impact on individuals [67].

Similarly, the IPSS has been shown to differentiate between cases and controls for prevalence of LUTS [51,57,62]. The discriminative validity of the B-SAQ has not been explored (table 5).

Construct Validity: Convergent/Divergent Validity

Convergent validity for the ICIQ-UI SF has been examined in three included studies (table 5). Significant agreement has been found between item 1 (assessing frequency of leakage) and item 2 (assessing amount of leakage) of the ICIQ-UI SF and the Bristol Female Lower Urinary Tract Symptoms (BFLUTS) questionnaire ($r = 0.29$, $p = .002$; $r = 0.86$, $p < .001$, respectively) [69], and the quality of life item with The Kings Health Questionnaire (KHQ; $r = 0.72$, $p < .001$) [69] and with item 6 of the SF-36 questionnaire ($r = 0.49$, $p < .001$) [68]. Item 3 (assessing impact of leakage on everyday life) has been found to have a weaker correlation with other questionnaires including The International Continence Society Male Short Form (ICSMaleSF) questionnaire and the BFLUTS ($r = 0.24$, $p = .23$; to $r = 0.58$, $p < .001$) [69]. The total ICIQ-UI SF scores showed significant agreement with most of the KHQ subscales [65].

There is evidence of significant associations between IPSS scores and scores on several other questionnaires including: the Madsen-Iversen ($r = 0.85$) [50]; The Maine Medical Assessment Programme ($r = 0.88$) [50]; The Boyarsky ($r = 0.93$) [50]; The Psychological General Well-Being Index (PGWBI; $r = 0.14$ to 0.41) [50]; The EuroQol Visual Analogue Scale (EQ-VAS; $r = -0.29$) [62], and The EuroQol Five Dimensions Questionnaire (EQ-5D; $r = -0.07$ to 0.36) [62]. The convergent/divergent validity for the IPSS has also been examined by comparing the 7 symptom scores with item 8, the quality of life score ($r = 0.72$ and $r = 0.82$; table 5) [57,62] and by comparing the same categories within the IPSS ($r > 0.33$) [58] and different categories within the IPSS ($r < 0.33$) [58]. These consistent findings of significant agreements between IPSS scores and alternative

questionnaires and between items within the same categories of the IPSS, demonstrate that the IPSS displays overall good convergent/divergent validity.

The convergent/divergent validity of the B-SAQ has been assessed by comparing the B-SAQ symptom scores with the KHQ symptom severity score (r = 0.46 to 0.54; table 5) [74]. Sahai et al. [73] also found high levels of agreement between B-SAQ individual symptom scores and KHQ individual symptoms in the symptom score domain (agreement rates: frequency 86%, urgency 85%, nocturia 84% and urinary incontinence 79%, $P < 0.001$). Comparisons have also been made between KHQ incontinence impact domain scores and the B-SAQ total symptom score scores (r = 0.79 to 0.81) [74] and the B-SAQ total bother scores (r = 0.81) [74]. These results highlight that the B-SAQ also has good convergent validity.

[insert table 5 here]

Critical Appraisal of Included Studies

Differentiation regarding the quality of included studies is illustrated in table 6. Three studies had a positive response to all of the relevant items [58,62,66] implying high quality. Hashim et al. [67] has the lowest percentage of positive responses compared to the rest of the studies. However, two other studies [63,69] had the same amount of "No" responses as Hashim et al. [67], implying not that these studies were of poor quality but that they had the poorest quality when judged against the other included studies.

[insert table 6 here]

Discussion

The current systematic review found three generic bladder health screening questionnaires that have been evaluated for their psychometric properties within a male and female population: the IPSS, the ICIQ-UI SF and the B-SAQ.

All three of the identified questionnaires had less than ten items, with the B-SAQ having fewer items than both the IPSS and the ICIQ-SF. For all the questionnaires there were

reports of low levels of missing data and high response rates for completing the questionnaire on a second occasion. However, there were some differences in findings for the IPSS in regards to participant understanding. A possible explanation for the differences is that the studies which report the IPSS to be complicated were conducted in non-English speaking samples with non-English versions of the IPSS; Turkey, Italy, Brazil and Argentina [52-54] whereas Barry et al. [51] who report the IPSS easy to understand conducted their research using the original English version of the IPSS with English speaking patients. Simple summation is the scoring system used for all three questionnaires. This review therefore concludes that the IPSS, ICIQ-UI SF and the B-SAQ are all quick and easy to complete questionnaires, potentially suitable for use by adult members of the public.

The IPSS had more validation and reliability studies and the largest sample sizes within studies, compared to the ICIQ-UI SF and the B-SAQ. The IPSS and the ICIQ-UI SF had more studies showing acceptable levels of test-retest reliability than the B-SAQ. However, the B-SAQ had only one assessment of test-retest reliability, which used a longer timescale between test and retest administrations compared to IPSS and ICIQ-SF studies. This may explain the differences between study findings. All three questionnaires show evidence of acceptable levels of internal consistency.

Only the IPSS has evidence of acceptable inter-rater reliability. It also has evidence of reliability between different forms of delivery. This lack of reliability findings for the ICIQ-UI SF and the B-SAQ does not illustrate a limitation of the tools as they were designed for self-completion and therefore does not devalue the psychometric strengths of these tools when used in the intended context. The additional validity testing of administration mode of the IPSS does however provide evidence that if the IPSS was used in a health promotion programme, there would be the option to use either a self-administration method or interview-administration method, and both would be equally reliable. One study did find that

the ICIQ-UI SF had better test-retest reliability when paper versus paper was compared, than when paper versus telephone was used. If the ICIQ-UI SF or the B-SAQ was to be used, self-administration would be required as there is a lack of evidence for acceptable levels of reliability associated with alternative delivery/completion methods.

Only the ICIQ-UI SF had an assessment of validity, using a gold standard criterion. The reason for absence of criterion-orientated validity assessments within the realm of urinary bladder health may be due to the lack of agreement within the literature as to what constitutes a gold standard measure for LUTS and/or urinary incontinence [76,77].

The ICIQ-UI SF and the IPSS both have good construct validity in regards to discriminative power however a greater variety of theoretical concepts associated with urinary bladder health has been explored with the ICIQ-UI SF, which focuses on urinary incontinence rather than LUTS, the focus of the IPSS. It is not possible to comment on the discriminative power of the B-SAQ as this was not assessed in the two relevant studies within this review.

All three questionnaires have evidence of convergent validity. The B-SAQ has the least validation assessment and has been compared with only one other questionnaire; the KHQ. The ICIQ-UI SF and the IPSS however have been compared with several other questionnaires with the majority of results showing significant levels of convergence. It is difficult however to determine which of these two questionnaires, the ICIQ-UI SF or IPSS, has the better convergent validity as authors have used different questionnaires in their assessments, other than the quality of life questionnaires. Nevertheless, the IPSS was the only questionnaire that had been assessed for divergent validity, thus making the evidence for convergent/divergent validity stronger than the ICIQ-UI SF and the B-SAQ.

The B-SAQ was last validated in 2014, however only two studies have validated the B-SAQ since its development in 2007. This lack of assessment also means that a limited

number of participants and types of populations have been used to test reliability and validity of the B-SAQ. There is less confidence therefore in the psychometric properties of the B-SAQ than the IPSS and ICIQ-UI SF, although more validation studies for all three questionnaires are desirable. The acceptability of all questionnaires appears to be good as the majority of participants either completed the full study or completed the questionnaire on two occasions. A more objective assessment of acceptability, via interview or survey methods, has not been undertaken for any of the screening questionnaires.

Many of the studies in the current review failed to provide adequate statistical results for their assessments of reliability and validity. A number of different measures of validity and reliability were also used making direct comparisons of psychometric properties impossible and it is recognised that this is a limitation of the review's analyses. The findings highlight the need for global definitions of each type of validity.

The results within this review demonstrate that the IPSS and ICIQ-UI SF are reliable and valid bladder health screening tools addressing different concepts; LUTS and urinary incontinence respectively. The UK National Health Service recognises the IPSS as a validated LUTS screening tool [78,79] and it has been proposed as a standard assessment for the screening of LUTS worldwide [52]. The ICIQ-UI SF is the recommended screening tool for UI in the NICE guidelines for the management of urinary incontinence in women [80]. The European Association of Urology class the ICIQ-UI SF as a "Category A" questionnaire for measuring UI symptoms [81]. The review by Hewison et al. [40] also found the ICIQ-UI SF to be suitable for use by time-pressured health practitioners due to the small number of items, low levels of missing data, and relevance to both community and clinical settings. The fact that both these questionnaires are recommended screening tools and are used in daily clinical practice gives more confidence regarding their validity and reliability for their different purposes.

Limitations and strengths

The applied stringent eligibility criteria may have limited the potential for identifying all available reliable and valid generic bladder health screening questionnaires. However, the objective was not only to identify screening questionnaires to assess their psychometric properties but also to identify those which would be applicable for use in a non-clinical context with the general public. Although most studies were conducted in specialist clinical contexts, the specific tools met the criterion that the questionnaire could be used by non-clinical staff and respondents. By considering applicability to a non-clinical population, including older community living adults, the recommendation for an appropriate bladder health screening questionnaire to be used as part of a community health promotion programme, can be made with more confidence.

The exclusion of non-English studies may be seen as a limitation however there was insufficient resources for translation and, given that the tools were developed in English it is considered unlikely that this restriction would have resulted in information relevant to UK practice being missed.

Literacy levels for the understanding and comprehension of the questionnaires were not considered in this review. One study in the current review found that education levels play a significant role in the ability to complete the IPSS questionnaire correctly. This review's findings also suggest that the translated versions of the IPSS may be more susceptible to the problems of lack of understanding than the original English versions and further investigation is required to confirm if linguistics is a possible contributor to poor understanding. It would also be useful to investigate if educational levels are independent influences on the understanding of the IPSS or if it is the interaction between education levels and the language versions of the questionnaire that has the greatest effect on this

understanding. Considering the evidence as it stands, health literacy should be a consideration when using the IPSS as part of a health promotion intervention.

The current review did not consider the development process of the questionnaires. Guidelines for the development and content validity evaluation of Patient Reported Outcome (PRO) instruments were established in 2009 [82] which highlight that concept evaluation interviews with patients are considered a fundamental cornerstone of these instruments. Patrick et al. [83] provide a gold standard step by step methodology to developing PRO instruments based on the 2009 guidelines which again place a large emphasis on target population input into the item generation of the instrument and also documentation provision of the process. The International Consultation on Incontinence 2017 guidelines also state [84, p.549]

> In addition to clinician input and literature review, questionnaire items must be generated from a patient perspective and include patient views. This is obtained through focus groups or one-to-one interviews …

Barry et al. [50] developed the AUASI (IPSS) before these guidelines were produced and this fundamental stage (concept elicitation) was not considered, stressing an important limitation of the IPSS. Nevertheless, the 2009 guidelines [83] suggest that existing questionnaires that did not consider this concept elicitation stage, may still be classed as being of good quality as long as trials have shown the questionnaire to have good levels of reliability. The current review found the IPSS to have good levels of reliability, widespread use and acceptance, which can be used as indicators of its acceptability.

The findings highlight that most of the psychometric assessments of the three identified questionnaires have been undertaken with people who have already identified they have LUTS and/or urinary incontinence and have sought healthcare. If these questionnaires are to be used for general screening and raising awareness of LUTS and/or urinary

incontinence amongst the general public, it is advisable that more psychometric assessments are conducted in non-clinical contexts.

Conclusion

The current review has highlighted that there are very few reliable and valid generic bladder health screening tools available. The reliability and validity of the IPSS and the ICIQ-UI SF have been more rigorously tested than the B-SAQ. Furthermore, the IPSS has stronger evidence for inter-rater reliability and convergent/divergent validity than the ICIQ-UI SF however they screen for different conditions. The ICIQ-UI SF only screens for UI and does not capture non-leakage urinary symptoms, meaning that potential opportunities could be lost to prevent the progression of urinary symptoms to urinary incontinence [2]. The IPSS provides assessment of storage and voiding LUTS without the screening of urinary incontinence. However urinary incontinence is a major contributor to avoidance of activity and social engagement in older adults [20-23], and is potentially amenable to supported self-management and bladder health promotion interventions. The B-SAQ covers both LUTS and urinary incontinence but focuses on certain type of LUTS (storage) which results in missed detection of important voiding symptoms.

In summary, the IPSS and the ICIQ-UI SF are generic bladder health screening questionnaires that have been evaluated for their psychometric properties within male and female adult populations in specialist and generic public health contexts. They are brief, easy and fast to complete, use simple scoring systems and are associated with low levels of missing data. They have both been subjected to the most rigorous psychometric testing in comparison to the B-SAQ. It is suggested that the IPSS be used when screening for LUTS and the ICIQ-UI SF be used when screening specifically for urinary incontinence. Both questionnaires are recommended as good options for screening for bladder health in a

community health promotion programme to tackle disability outcomes associated with LUTS.

Acknowledgements: This research was funded on a PhD studentship from Glasgow Caledonian University. The funding source had no specific involvement in this research.

The authors report no conflict of interest.

References

1. Abrams P, Cardozo L, Fall M, et al. Standardisation Sub-Committee of the International Continence Society. The standardisation of terminology in lower urinary tract function: Report from the standardisation sub-committee of the international continence society. Urology. 2003 Jan;61(1):37-49.

2. Irwin DE, Kopp ZS, Agatep Bet al. Worldwide prevalence estimates of lower urinary tract symptoms, overactive bladder, urinary incontinence and bladder outlet obstruction. BJU Int. 2011 Oct;108(7):1132-8.

3. Haylen BT, de Ridder D, Freeman RM, et al. An international urogynecological association (IUGA)/International continence society (ICS) joint report on the terminology for female pelvic floor dysfunction. Int Urogynecol J. 2010 Jan;21(1):5-26.

4. Coyne KS, Sexton CC, Kopp ZS, et al. The impact of overactive bladder on mental health, work productivity and health-related quality of life in the UK and Sweden: Results from EpiLUTS. BJU Int. 2011 Nov;108(9):1459-71.

5. Hajjar RR. Psychosocial impact of urinary incontinence in the elderly population. Clin Geriatr Med. 2004;20(3):553-64.

6. Chiverton PA, Wells TJ, Brink CA, et al. Psychological factors associated with urinary incontinence. Clinical Nurse Specialist. 1996;10(5):229-33.

7. Patel DP, Elliott SP, Stoffel JT, et al. Patient reported outcomes measures in neurogenic bladder and bowel: A systematic review of the current literature. Neurourol Urodyn. 2016;35:8-14.

8. Voegeli D. Incontinence-associated dermatitis: New insights into an old problem. Br J Nurs. 2016 Mar 10-23;25(5):256, 258, 260-2.

9. Bedretdinova D, Fritel X, Zins M, et al.The effect of urinary incontinence on health-related quality of life: Is it similar in men and women? Urology. 2016 May;91:83-9.

10. Minassian VA, Sun H, Yan XS, et al. The interaction of stress and urgency urinary incontinence and its effect on quality of life. Int Urogynecol J. 2015 Feb;26(2):269-76.

11. Vandoninck V, Bemelmans BL, Mazzetta C, et al. UREPIK study group. The prevalence of urinary incontinence in community-dwelling married women: A matter of definition. BJU Int. 2004 Dec;94(9):1291-5.

12. Gosch M, Talasz H, Nicholas JA, et al. Urinary incontinence and poor functional status in fragility fracture patients: An underrecognized and underappreciated association. Arch Orthop Trauma Surg. 2015;135(1):59-67.

13. Brown JS, Vittinghoff E, Wyman JF, et al/ Urinary incontinence: Does it increase risk for falls and fractures? study of osteoporotic fractures research group. J Am Geriatr Soc. 2000 Jul;48(7):721-5.

14. Thom DH, Haan MN, Van Den Eeden SK. Medically recognized urinary incontinence and risks of hospitalization, nursing home admission and mortality. Age Ageing. 1997 Sep;26(5):367-74.

15. Bertera EM. Depression in older Americans with urinary incontinence (UI) the relationship with activities of daily living (ADL) and avoidance behaviors. Journal of Gerontological Social Work. 2003;39(4):39-53.

16. Thorn DH, Brown JS. Reproductive and hormonal risk factors for urinary incontinence in later life: A review of the clinical and epidemiologic literature. J Am Geriatr Soc. 1998;46(11):1411-7.

17. Hu TW, Wagner TH, Bentkover JD, et al. Costs of urinary incontinence and overactive bladder in the united states: A comparative study. Urology. 2004 Mar;63(3):461-5.

18. Offermans MP, Du Moulin MF, Hamers JP, et al. Prevalence of urinary incontinence and associated risk factors in nursing home residents: A systematic review. Neurourol Urodyn. 2009;28(4):288-94.

19. Coll-Planas L, Denkinger MD, Nikolaus T. Relationship of urinary incontinence and late-life disability: Implications for clinical work and research in geriatrics. Zeitschrift für Gerontologie und Geriatrie. 2008;41(4):283-90.

20. Stadnicka G, Lepecka-Klusek C, Pilewska-Kozak AB, et al. Psychosocial problems of women with stress urinary incontinence. Annals of Agricultural and Environmental Medicine. 2015;22(3)

21. Monz B, Pons ME, Hampel C, Hunskaar S, et al. Papanicolaou S. Patient-reported impact of urinary incontinence—Results from treatment seeking women in 14 European countries. Maturitas 2018/03;52:24-34.

22. Nygaard I, Girts T, Fultz NH, Kinchen K, et al. Is urinary incontinence a barrier to exercise in women? Obstet Gynecol. 2005 Aug;106(2):307-14.

23. Brown WJ, Miller YD. Too wet to exercise? leaking urine as a barrier to physical activity in women. Journal of Science and Medicine in Sport 2018/03;4(4):373-8.

24. Irwin DE, Milsom I, Hunskaar S, et al. Population-based survey of urinary incontinence, overactive bladder, and other lower urinary tract symptoms in five countries: Results of the EPIC study. Eur Urol. 2006 Dec;50(6):1306,14; discussion 1314-5.

25. McGrother CW, Donaldson MM, Shaw C, et al. Storage symptoms of the bladder: Prevalence, incidence and need for services in the UK. BJU Int. 2004 Apr;93(6):763-9.

26. Crestodina LR. Assessment and management of urinary incontinence in the elderly male. Nurse Pract. 2007 Sep;32(9):26,34; 34-5.

27. Anger JT, Saigal CS, Litwin MS, Urologic Diseases of America Project. The prevalence of urinary incontinence among community dwelling adult women: Results from the national health and nutrition examination survey. J Urol. 2006 Feb;175(2):601-4.

28. Gibson W, Wagg A. New horizons: Urinary incontinence in older people. Age Ageing. 2014 Mar;43(2):157-63.

29. Tannenbaum C, Agnew R, Benedetti A, et al. Effectiveness of continence promotion for older women via community organisations: A cluster randomised trial. BMJ Open [Internet]. 2013 British Medical Journal Publishing Group;3(12)

30. Elstad EA, Taubenberger SP, Botelho EM, et al. Beyond incontinence: The stigma of other urinary symptoms. J Adv Nurs. 2010 Nov;66(11):2460-70.

31. Horrocks S, Somerset M, Stoddart H, et al. What prevents older people from seeking treatment for urinary incontinence? A qualitative exploration of barriers to the use of community continence services. Fam Pract. 2004 Dec;21(6):689-96.

32. Dugan E, Roberts CP, Cohen SJ, et al. Why older community-dwelling adults do not discuss urinary incontinence with their primary care physicians. J Am Geriatr Soc. 2001 Apr;49(4):462-5.

33. Teunissen D, van Weel C, Lagro-Janssen T. Urinary incontinence in older people living in the community: Examining help-seeking behaviour. Br J Gen Pract. 2005 Oct;55(519):776-82.

34. Sievert KD, Amend B, Toomey PA, et al. Can we prevent incontinence? ICI-RS 2011. Neurourol Urodyn. 2012 Mar;31(3):390-9.

35. Wallner LP, Porten S, Meenan RT, et al. Prevalence and severity of undiagnosed urinary incontinence in women. Am J Med. 2009 11;122(11):1037-42.

36. Andrade AD, Anam R, Karanam C,et al. An overactive bladder online self-management program with Embedded Avatars: A randomized controlled trial of efficacy. Urology. 2015 3;85(3):561-7.

37. Wyman JF, Burgio KL, Newman DK. Practical aspects of lifestyle modifications and behavioural interventions in the treatment of overactive bladder and urgency urinary incontinence. Int J Clin Pract. 2009 Aug;63(8):1177-91.

38. Parsons JK, Wilt TJ, Wang PY, et al. Osteoporotic Fractures in Men Research Group. Progression of lower urinary tract symptoms in older men: A community based study. J Urol. 2010 May;183(5):1915-20.

39. Rohrmann S, Katzke V, Kaaks R. Prevalence and progression of lower urinary tract symptoms in an aging population. Urology 2018/03;95:158-63.

40. Hewison A, McCaughan D, Watt I. An evaluative review of questionnaires recommended for the assessment of quality of life and symptom severity in women with urinary incontinence. J Clin Nurs. 2014;23(21-22):2998-3011.

41. Bowling A. Techniques of questionnaire design. Maidenhead (UK): Open University Press; 2005.

42. Naughton MJ, Donovan J, Badia X, et al. Symptom severity and QOL scales for urinary incontinence. Gastroenterology. 2004;126:S114-23.

43. McColl, E, Jacoby, A, Thomas, L, et al. Design and use of questionnaires: a review of best practice applicable to surveys of health service staff and patients. Core Research; 2001.

44. Moher D, Shamseer L, Clarke M, et al. Preferred reporting items for systematic review and meta-analysis protocols (PRISMA-P) 2015 statement. Systematic Reviews. 2015 01/01;4(1):1.

45. Bellet RN, Adams L, Morris NR. The 6-minute walk test in outpatient cardiac rehabilitation: Validity, reliability and responsiveness—a systematic review. Physiotherapy. 2012;98(4):277-86.

46. Kersten P, Czuba K, McPherson K, et al. A systematic review of evidence for the psychometric properties of the strengths and difficulties questionnaire. Int J Behavioral Development. 2015:0165025415570647.

47. Helmerhorst HHJ, Brage S, Warren J, et al. A systematic review of reliability and objective criterion-related validity of physical activity questionnaires. Int J Behavioral Nutrition and Physical Activity. 2012;9(1):1.

48. Lubans DR, Hesketh K, Cliff D, et al. A systematic review of the validity and reliability of sedentary behaviour measures used with children and adolescents. Obesity reviews. 2011;12(10):781-99.

49. Lohr KN. Assessing health status and quality-of-life instruments: Attributes and review criteria. Quality of Life Research. 2002;11(3):193-205.

50. Barry MJ, Fowler FJ,Jr, O'Leary MP, et al. Correlation of the American urological association symptom index with self-administered versions of the Madsen-Iversen, Boyarsky and Maine medical assessment program symptom indexes. Measurement committee of the American urological association. J Urol. 1992 Nov;148(5):1558,63; discussion 1564.

51. Barry MJ, Fowler FJ,Jr, O'Leary MP, et al. The american urological association symptom index for benign prostatic hyperplasia. the measurement committee of the american urological association. J Urol. 1992 Nov;148(5):1549,57; discussion 1564.

52. Cam K, Akman Y, Cicekci B, et al. Mode of administration of international prostate symptom score in patients with lower urinary tract symptoms: Physician vs self. Prostate cancer and prostatic diseases. 2004;7(1):41-4.

53. Russo F, Di Pasquale B, Romano G, et al. International prostate symptom score: Comparison of doctor and patient. Arch Ital Urol Androl. 1998 Jun;70(3 Suppl):15-24.

54. Netto NR, de Lima ML, de Andrade EFM, et al. Latin american study on patient acceptance of the international prostate symptom score (IPSS) in the evaluation of symptomatic benign prostatic hyperplasia. Urology [Internet]. 1997 January 1997;49(1):46-9.

55. Johnson TV, Schoenberg ED, Abbasi A, et al. Assessment of the performance of the american urological association symptom score in 2 distinct patient populations. J Urol. 2009 Jan;181(1):230-7.

56. Barry MJ, Fowler FJ, Chang Y, et al. The American urological association symptom index: Does mode of administration affect its psychometric properties? J Urol. 1995 1995;154(3):1056-9.

57. Hammad FT, Kaya MA. Development and validation of an Arabic version of the international prostate symptom score. BJU Int. 2010 May;105(10):1434-8.

58. Okamura K, Nojiri Y, Osuga Y, et al. Psychometric analysis of international prostate symptom score for female lower urinary tract symptoms. Urology. 2009 Jun;73(6):1199-202.

59. Quek KF, Chua CB, Razack AH, et al. Construction of the mandarin version of the international prostate symptom score inventory in assessing lower urinary tract symptoms in a Malaysian population. Int J Urol. 2005 Jan;12(1):39-45.

60. O'Connor RC, Bales GT, Avila D, et al. Variability of the international prostate symptom score in men with lower urinary tract symptoms. Scand J Urol Nephrol. 2003;37(1):35-7.

61. Garcia-Losa M, Unda M, Badia X, et al. Effect of mode of administration on I–PSS scores in a large BPH patient population. Eur Urol. 2001;40(4):451-7.

62. Badia X, Garcia-Losa M, Dal-Re R, et al. Validation of a harmonized spanish version of the IPSS: Evidence of equivalence with the original American scale. international prostate symptom score. Urology. 1998 Oct;52(4):614-20.

63. El Din KE, Koch W, De Wildt M, et al. Reliability of the international prostate symptom score in the assessment of patients with lower urinary tract symptoms and/or benign prostatic hyperplasia. J Urol. 1996;155(6):1959-64.

64. Lim R, Liong ML, Lau YK, et al. Validity, reliability, and responsiveness of the ICIQ-UI SF and ICIQ-LUTSqol in the Malaysian population. Neurourol Urodyn. 2017 Feb;36(2):438-42.

65. Gotoh M, Homma Y, Funahashi Y, et al. Psychometric validation of the japanese version of the international consultation on incontinence questionnaire-short form. Int J Urol. 2009 03;16(3):303-6.

66. Rotar M, Trsinar B, Kisner K, et al. Correlations between the ICIQ-UI short form and urodynamic diagnosis. Neurourol Urodyn. 2009;28(6):501-5.

67. Hashim H, Avery K, Mourad M, et al. The arabic ICIQ- UI SF: An alternative language version of the english ICIQ- UI SF. Neurourol Urodyn. 2006;25(3):277-82.

68. Tubaro A, Zattoni F, Prezioso D, et al. W, FLOW Study Grp. Italian validation of the international consultation on incontinence questionnaires. BJU Int. 2006 JAN 2006;97(1):101-8.

69. Avery K, Donovan J, Peters TJ, et al. ICIQ: A brief and robust measure for evaluating the symptoms and impact of urinary incontinence. Neurourol Urodyn. 2004 2004;23(4):322-30.

70. Kurzawa Z, Sutherland JM, Crump T, et al. Measuring quality of life in patients with stress urinary incontinence: is the ICIQ-UI-SF adequate? Qual Life Res. 2018 08;27(8):2189-2194.

71.MaryHeck G., Krief H., Akrour R., et al. Screening for urinary incontinence in acute care for elders unit: comparative performance analysis of Katz's ADL and ICIQ-UI-SF. European Geriatric Medicine. 2018;9(5):579-588.

72. Uren AD, Cotterill N, Parke SE, et al. Psychometric equivalence of electronic and telephone completion of the ICIQ modules. Neurourol Urodyn. 2017;36(5):1342-1349.

73. Sahai A, Dowson C, Cortes E, et al. Validation of the bladder control self-assessment questionnaire (B-SAQ) in men. BJU Int. 2014;113(5):783-8.

74. Basra R, Artibani W, Cardozo L, et al. Design and validation of a new screening instrument for lower urinary tract dysfunction: The bladder control self-assessment questionnaire (B-SAQ). Eur Urol. 2007;52(1):230-8.

75. Tarrant C, Baker R, Colman AM, et al. The prostate care questionnaire for patients (PCQ-P): Reliability, validity and acceptability. BMC Health Serv Res. 2009 Nov 4;9:199,6963-9-199.

76. Nitti VW. Pressure flow urodynamic studies: The gold standard for diagnosing bladder outlet obstruction. Rev Urol. 2005;7 Suppl 6:S14-21.

77. Martin JL, Williams KS, Abrams KR, et al. Systematic review and evaluation of methods of assessing urinary incontinence. Health Technol Assess. 2006 Feb;10(6):1,132, iii-iv.

78. Pickard M, Hilmy M [Internet]. UR03 Lower Urinary Tract Symptoms in Men. [updated 2016 July 14; cited 2017 March 20]. Available from: http://www.valeofyorkccg.nhs.uk/rss/data/uploads/urology/ur03-lower-tract-symptoms-luts-in-men-published.pdf

79. Royal United Hospital Bath NHS Trust [Internet]. International Prostate Symptom Score (IPSS). [updated 2017 February; cited 2017 March 20]. Available from: http://www.ruh.nhs.uk/patients/Urology/documents/patient_leaflets/Form_IPSS.pdf

80. NICE: National Institute for Health and Clinical Excellence [Internet]. Urinary Incontinence in Women: Management; Clinical guideline CG171. [updated 2015 November; cited 2017 March 30]. Available from: https://www.nice.org.uk/guidance/CG171/chapter/1-Recommendations

81. Lucas MG, Bedretdinova D, Berghmans LC, et al. Guidelines on Urinary Incontinence:urinary incontinence - partial update march 2015. European Association of Urology; 2015;9;210-236

82. FDA U.S. Food and Drug Administration (FDA) [Internet]. FDA U.S. food and drug administration patient-reported outcome measures: Use in medical product development to support labeling claims. [updated 2009 December; cited 2018 February 21]. Available from: https://www.fda.gov/downloads/drugs/guidances/ucm193282.pdf

83. Patrick DL, Burke LB, Gwaltney CJ, et al. Content validity establishing and reporting the evidence in newly developed patient-reported outcomes (PRO) instruments for medical product evaluation: ISPOR PRO good research practices task force report: Part 2 assessing respondent understanding. Value Health. 2011. Dec;14(8):978-88

84. Diaz DC, Bosch R, Costantini E et al. Patient-reported outcome assessment. In: Abrams P, Cardazo L, Wagg A, et al., editors. Incontinence: 6th International Consultation on Incontinence, Tokyo, September 2016. 6th ed. Vol. 1. ICUD ICS; 2017. p. 541-598.

FIGURE CAPTIONS:

Figure 1: PRISMA flow diagram of screening and selection process.

table 1:	Psychometric Definitions Used to Operationalise Each Psychometric Measure
	Definition
Reliability	
Test-retest Reliability	Test-retest reliability refers to the likelihood of the same instrument obtaining similar/stable results when administered to the same group of heterogeneous subjects at different time-points [44]. When assessing test-retest reliability it is important to consider the timescales between the different tests/administrations as variables are likely to change over time. If the timescale is too long, the changes within variables will affect the results of the reliability assessment. Alternatively, enough time must be provided to eliminate/reduce the possibility of practice effects, which could exaggerate the agreement found between tests/administrations [45].
Internal Consistency	Internal consistency reliability refers to the homogeneity of the instrument (that is, the degree of interrelatedness among items), whereby responses to items measuring the same variable should yield similar or consistent results from only one time-point, indicating good reliability [46]. Internal consistency may measure the level of association between each item (Cronbach's Alpha) or the correlation between one half of an individual's responses with the other half of the same person's responses (split half reliability; [47].
Inter-rater Reliability	Inter-rater reliability assessment compares the level of agreement between two or more raters in their observations/assessment of a variable [45,48].
Validity	
Content Validity	Content validity may include either logical validity or face validity tests. Logical validity refers to the considered opinion of expert judges as to the extent to which the items of a questionnaire comprehensively represent the concept of interest [47,49]. Face validity is the extent to which a questionnaire is perceived as covering the concept it aims to measure. Content validity which is not quantified by statistics is a subjective type of validity, therefore being a less sophisticated form of validity than criterion-oriented validity and construct validity, which are objective forms of validity [50]. However, testing for content validity is still important for making an assessment of the applicability of a questionnaire [39,45].
Criterion Validity	Criterion-orientated validity is a measure of how well one instrument compares with another instrument or predictor, called a criterion. Concurrent and predictive validity are two components of criterion validity [46].
Concurrent Validity	Concurrent validity is determined by assessing the correlation between the questionnaire and another, more objective measure of the same domain, often acknowledged as the gold standard method, when administered simultaneously [40,45,46]. An example of a "gold standard" criterion for assessing UI is urodynamic testing [51].
Predictive Validity	Predictive validity is assessed by measuring the correlation of an instrument with a gold standard criterion that will be available in the future [54].
Construct Validity	Construct validity which includes discriminative, convergent and divergent (discriminant) validity [46] is a theoretical measure of how meaningful (truthful) a scale or instrument is when in practical use [45].
Construct: Discriminative Validity	Discriminative validity is the measurement of an instrument's ability to discriminate between two distinct groups [46].
Construct: Convergent Validity	Convergent validity is the degree to which the scores of one instrument relates to scores of another instrument which is expected to be a measure of the same realm [40,52]. In the assessment of UI, this is often a test of correlation with another condition specific questionnaire.

Construct: Divergent Validity	Divergent (discriminant) validity is the degree to which the scores of one instrument's scores do not correlate with scores of another instrument which measures disparate constructs, i.e. the ability to discriminate between different constructs [46,52].

table 2: Screening Questionnaire Characteristics

Questionnaire	Type of bladder problem screened	Number of items	Scoring system	Time taken to complete	Missing data	Dropout/response rate
ICIQ-UI SF	UI	4 items (Avery, 2004) [69]	The ICIQ-UI Short Form comprises four items, three of which are scored and summed to yield the total score (score 0–21). These questions assess the frequency of leakage (score 0–5), the amount of leakage (score 0–6), and the impact of incontinence on the quality of life (score 0–10). The fourth unscored question is self-diagnostic about the perceived causes of incontinence. The higher the score the greater the severity. (Avery, 2004; Hashim 2006; Rotar 2009; Gotoh, 2009; Tubaro, 2006; Lim, 2017, Kurzawa, 2018; Mary-Heck, 2018; Uren, 2017).		Levels of missing data; (mean 1.6% range, <1% to 2%: Avery, 2004); (<1%) for all items: Hashim 2006; (<1%): Lim, 2017. 89% of participants completed all items. (Avery, 2004) 1 question unanswered by 1 person at T1 but was answered at T2 (Rotar, 2009) Missing responses to any one question = 1.02% (Kurzawa, 2018)	Lowest response rate = 60%, highest response rate = 96%. (Avery, 2004).
B-SAQ	LUTS and UI	8 items (Basra, 2007). [74]	The B-SAQ comprises eight items, four items assess symptoms and four assess bother. Individual symptom and bother scores are summated to give an overall symptom and bother score. A score interpretation table grades the severity of symptoms and bother. The questionnaire recommends seeking medical attention if the symptom score is ≥4 and states that the patient may benefit from help if the bother score is ≥1. A separate statement in bold at the bottom of the questionnaire specifically identifies "red flags" and warns that those patients with haematuria, voiding difficulty or pain on passing urine should consult their doctor immediately. (Basra, 2007; Sahai, 2014).	A total of 98% (Sahai, 2014); 89% (Basra, 2007) of participants completed all items correctly in <5 min. (Sahai, 2014; Basra, 2007).	A total of 98% (Sahai, 2014); 89% (Basra, 2007) of participants completed all items correctly (Sahai, 2014; Basra, 2007).	89% correctly completed and returned the B-SAQ (Basra, 2007)
IPSS/AUASI	LUTS. LUTS (O'Connor, 2003).	7 item; IPSS has additional item measuring impact of symptoms	The IPSS comprises seven items that measure symptom frequency and severity, each of which is rated from 0 (not at all) to 5 (almost always), plus one disease specific quality of life question. The total score is the sum of items 1–7 (range: 0–35) according to the severity of symptoms. (Quek 2005). Symptoms are considered mild for scores between 0 and 7, moderate for		Questions were clear to 95% of participants. (Barry 1992b). 72% understood all	87% of participants returned questionnaire at time 2. (Barry 1992b). 60% of participants returned

	(Badia, 1998) [62]	scores between 8 and 19, and severe for scores between 20 and 35. (Badia, 1998; Johnson, 2009).	AUASI questions (Johnson, 2009). 54% of patients were unable to complete the IPSS in full for lack of understanding Most of these patients (49%) had a lower educational level (assessed by attainment level ranging from elementary school level to university degree level)(Cam, 2004). 89% completed all questions. Missing responses to individual items ≤7.5%. (Okamura 2009).	the questionnaires. (O'Connor, 2003). 91% of participants completed the full study. (Johnson, 2009). 97.6% of patients attended visit 2 (Garcia-Losa, 2001). 96% of patients attended for follow-up (Hammad, 2010).

ICIQ-UI SF = International Consultation on Incontinence, Urinary Incontinence Short Form; IPSS = International Prostate Symptom Score; B-SAQ – Bladder Control Self-Assessment Questionnaire; LUTS = Lower Urinary Tract Symptom; UI = urinary incontinence; AUASI American Urological Association Symptom Index.

table 3: Study Characteristics

				Participant Characteristics			
Reference	**Screening Questionnaire/ version**	**Sample size**	**Mode of administration**	**Age**	**Gender**	**Health status**	**Context of Screening**
Avery, 2004 [69]	ICIQ-UI SF/ English	1306 (from all studies) 469, 223, 246, 57, 206, 105. (used 6 samples as had multiple studies)	SA	Mean age (from all studies) ranged from 50.1 years to 66.6 years.	Both - 800 females/ 506 males in total. Female to male ratio = (61% - 39%)	Differing levels of urinary symptoms. No more information provided.	Community-based samples and urology clinic attendees used. Specialist services – clinicians.
Basra, 2007 [74]	B-SAQ/ English	329	SA	Mean age 47 years (Range 18-83 years).	All female.	All women were attending a general gynaecology or urogynaecology clinic. Clinical diagnosis from the doctors' evaluation sheet defined 31% of participants were asymptomatic controls.	Screening took place during consultation at a gynaecology or urogynaecology outpatient clinic. Specialist Services – clinicians.
Sahai, 2014 [73]	B-SAQ/ English	211	SA	Cases mean age 62 years, controls mean age 41 years.	All male.	Male patients attending a busy urology or dedicated stone outpatient clinic. Cases confirmed by Urologist if LUTS was present in a consultation.	Took place at outpatient urology clinic. Specialist services.
Barry, 1992(a) [50]	AUA SI/ English	135	SA	Not provided. But control group were aged between 18 and 55 years. Details provided in Barry 1992 b – see below.	All male.	76 men with clinically defined BPH, 59 men being evaluated for non-urological complaints but with no history of urinary or prostate problems.	3 urology practices. Specialist services.
Barry 1992(b) [51]	AUA SI/ English	318 (210 BPH patients and 108 controls.)	SA	1st validation study BPH patients mean age = 64 ± 9 years, range 44 to 82 years, control subjects mean age 39 ± 9 years, range 21 to 55 years). 2nd validation study	All male.	Men with BPH and controls without.	3 urology practices. Specialist services.

				= 107 BPH patients and 49 controls. Age distribution almost identical to 1st validation study. Final phase of validation mean age is 66 years.			
Barry, 1995 [56]	AUA SI/ English	124	SA versus physician administration	Group 1 mean age = 68.7 (sd 10.2), Group 2 mean age = 68.7 (SD 7.5) group 3 mean age = 67.4 (SD 7.8).	All male.	41 visually impaired or illiterate men and 83 able men. All had history of symptomatic BPH.	Screening took place at 10 specialist clinical practices.
O'Connor, 2003 [60]	IPSS/ English	210	SA	Mean age = 67 years (range 41-94).	All male.	Men with LUTS who were either on a stable medical regimen to treat their symptoms or were receiving no therapy.	Screening took place in urologist office. Specialist services.
Johnson, 2009 [55]	AUA SI/ English	407	SA and interview Administration.	Mean age = 59.13 years.SD = 13.89.	Males (although not stated)	Consecutive patients from the urology clinics.	Took place in urology clinics. Specialist services.
El Din, 1996 [63]	IPSS/ English	71	SA at time 1. SA and physician administration at time 2.	Mean age = 63 years (sd = +-9) range = 44 to 83 years.	All male.	Men with BPH and/or LUTS who had been referred to a prostate centre.	Prostate centres. Specialist services – physician.
Cam, 2004 [52]	IPSS/ Turkish	150	SA versus physician administration.	58±8 y (range 50–79 y)	Not stated	Patients who had been admitted to an outpatient department due to lower urinary tract symptoms were recruited.	Outpatient urology clinic. Specialist services.
Garcia-Losa, 2001 [61]	IPSS/ Spanish	946	SA versus interview administration.	Mean age 65.57 (7.80).	Males (does not state this but inclusion criteria were patients with BPH).	Patients over 50 years of age diagnosed as having BPH according to clinical criteria, with any duration of disease evolution and severity, were recruited consecutively. Clinicians rated health status related to BPH as good or quite good in 52% of the patients in the sample as a whole.	52 urology clinics throughout Spain. Specialist services.
Okamura, 2009 [58]	IPSS/Japanese	1620	SA	Sample A women mean age 70.1 ± 8.4, Sample A men mean age 71.1 ± 9.1. Sample B women mean age 67.5 _±9.1. Sample B men mean age 66.7 ± 8.5.	Both - 746 women, 874 men. Female to male ratio = 46% - 54%).	(sample A) was collected from outpatients consulting hospital doctors and the other (sample B) was from outpatients consulting general practitioners for various chronic medical problems.	Consulting hospital doctor or consulting a general practitioner. Non-specialist services but not randomised as seeking help.

Rotar 2009 [66]	ICIQ-UI-SF/ Slovene	120 plus 70 plus 69 plus 54. (313)	SA	Validation process – baseline – 120 participants, 40% men, mean age 66years (±12). Stability, 70 participants, 40% men, mean age 69 years (±11). Responsiveness, 90 participants, 41% men, mean age 59 years (±13). Urodynamic, 54 participants, 6% men, mean age 60 years (±13).	Both. – Male to female ratio % = study 1 40:60. Study 2 40:60. Study 3 41:59. Study 4 6:94. Inclusive male to female ratio = 40% - 60%.	The baseline sample of patients was composed of patients with LUTS who attended urology or gynaecology outpatient clinics and elderly institutionalized patients. All the participants had satisfactory cognitive function. Stability was tested in a group of 70 patients that consisted of urology or gynaecology outpatient clinic attendants and elderly institutionalized with various levels of LUTS.	Included urology and gynaecology outpatient clinic attendees, elderly institutionalized patients and a community sample. Specialist services and non-specialist services.
Badia, 1998 [62]	IPSS/ Spanish version	127 (59 patients and 68 controls).	SA	Patients mean age 65.8 years ± 8.3. Controls mean age 36.8 years ± 10.2.	Males (although not stated but recruited patients with BPH).	78% of patients had moderate or severe urinary symptoms. Exclusion criteria for controls was any current or past clinical diagnosis related to urinary problems or BPH, or history of renal, vesical, or other chronic urinary disease.	Screening took place in 6 urologic clinics. Patients with BPH and control subjects without BPH. Specialist services – clinicians.
Gotoh, 2009 [65]	ICIQ-UI SF/ Japanese version	122 patients (Study 1)	SA	median age: 62 years; range: 53–70 years	Both - female patients comprised 83.6% of the study group.	All participants had UI.	Not stated where participants were recruited from.
Hammad 2010 [57]	IPSS/ Arabic version	139 men (76 patients and 63 normal subjects)	SA	Mean age control group 32.6 years (8.1). Mean age patients 61.4 years (9.1).	All Male.	76 patients with urinary symptoms due to BPH and in 63 control subjects without BPH; 25 patients had transurethral resection of prostate (TURP) whereas the remaining 51 patients were treated with terazosin.	Took place at urology clinic visit. Specialist services.
Quek, 2005 [59]	IPSS/ Mandarin version	68 participants (39 BPH group, 29 Control group).	SA	BPH group (n = 39; mean age: 70.64 ± 8.51 years), control group (n = 29; mean age: 63.37 ± 12.96 years).	All Male.	Patient group reported BPH. Control group were patients with renal stones.	Recruited from urology clinic and ward. Specialist services.
Tubaro 2006 [68]	ICIQ-UI SF/ Italian	103 participants.	SA	Cases and controls were enrolled in homogeneous age classes (≤ 50 or > 50 years). (44 ≤ 50 years, 45 > 50	All females	Cases: female patients who had been having LUTS for ≥ 3 months. Controls: healthy women who had had no LUTS for ≥ 3 months.	Took place in 4 Italian urology centres. Specialist services.

				years).			
Hashim, 2006 [67]	ICIQ-UI SF/ Arabic	For validity testing = 131 participants. For reliability testing = 102 of the participants	Not stated	For validity testing = 131 participants – 44 male 87 female, mean age 37.8, range 18-73 years. For reliability testing = 102 participants – 34 male 68 female, mean age 37.7 years, range 17-73 years).	Both - Validity testing 34% male. Reliability testing 33% male.	Patients attending urology clinics with varying levels of UI.	Urology outpatient clinics. Specialist services.
Lim, 2017 [64]	ICIQ-UI SF/ English, Malay and Chinese	248 participants (139 Stress UI patients, 145 healthy control group)	SA	Participants with Stress UI Mean age 52.2 ± 8.61 years. (Range 31 – 78 years). Control group Mean age 45.7 ± 9.59 years. (Range 23 – 66 years)	All female.	Cases: Female patients diagnosed with Stress UI. Control group: female volunteers who were continent.	Urology outpatient clinic in Malaysia. Specialist services.
Kurzawa, 2018 [70]	ICIQ-UI SF/ English Version	177 participants.	SA	Mean age 68.86 ± 8.71 years.	All male.	Participants with moderate to high symptom burden. Only patients undergoing initial treatment for SUI were contacted. Pre-operative patients.	Urology outpatient clinic in Canada. Specialist services.
Mary-Heck, 2018 [71]	ICIQ-UI SF/ English Version	294 participants.	Nurse	Mean age = 86.2 ± 6.5 years.	76.5% female.	Vulnerable, elderly patients (at least 75 years old) with at least 1 geriatric syndrome and with spontaneous micturition.	Acute Care for Elders (ACE) unit of hospital in Switzerland.
Uren, 2017 [72]	ICIQ-UI SF/ English Version	491 participants.	SA vs Telephone administration.	Not stated.	Males and females. Ratio not stated.	Patients attending urology clinics with complaints of LUTS.	Urology outpatient clinic in England. Specialist services.

Key: B-SAQ = Bladder Control Self-Assessment Questionnaire; AUA SI = American Urological Association Symptom Index; IPSS = International Prostate Symptom Score; SA = self-administration; ICIQ-UI SF = International Consultation on Incontinence-Urinary Incontinence Short Form; UI = urinary incontinence; LUTS = Lower Urinary Tract Symptoms; BPH = benign prostatic hyperplasia; specialist services = secondary care; non specialist services = primary care.

table 4: Reliability Results for Included Studies

Reference	Screening Questionnaire	Test-retest Period	Test-retest Result	Inter-rater reliability result	Internal Consistency Result
Lim, 2017 [64]	ICIQ-UI SF	1 week	ICC = 0.95 (English version) ICC = 0.91 (Chinese version) ICC = 0.96 (Malay version)	NR	Cα = 0.60 (English version) Cα = 0.61 (Chinese version) Cα = 0.76 (Malay version)
Avery 2004 [69]	ICIQ-UI SF	2 weeks	k = 0.74 (p<0.001)	NR	Cα = 0.92
Gotoh, 2009 [65]	ICIQ-UI SF	2 weeks	k = 0.61 for item 1 k = 0.62 for item 2 ICC = 0.90 for item 3 ICC = 0.91 for total score	NR	Cα = 0.78
Hashim, 2006 [67]	ICIQ-UI SF	2 weeks	k = 0.85 (p< 0.0001)	NR	Cα = 0.71
Rotar, 2009 [66]	ICIQ-UI SF	2 weeks	k = 0.99 for item 1 k = 0.98 for item 2 k = 0.95 for item 3	NR	Cα = 0.81
Kurzawa, 2018 [70]	ICIQ-UI SF	NR	NR	NR	Cα = 0.63 for all 3 items. Cα = 0.67 with deletion of 3rd item.
Uren, 2017 [72]	ICIQ-UI SF	Paper: 20 minutes Telephone: 1 week	Paper vs paper: k = 0.74 for item 1 k = 0.88 for item 2 k = 0.52 for item 3 Paper vs Telephone: k = 0.62 for item 1 k = 0.75 for item 2 k = 0.51 for item 3	NR	NR
Tubaro, 2006 [68]	ICIQ-UI SF	1 week	ICC = 0.93	NR	Cα = 0.90
Barry 1992 [50]	AUA SI	1 week	r = 0.92	NR	Cα = 0.86
Barry, 1995 [56]	AUA SI	1 week	ICC = 0.76 to 0.82 (range between groups)	No significant difference between raters. *p* = 0.48	Cα = 0.75 to 0.87 (range between time and group).
El Din, 1996 [63]	IPSS	8 weeks	r = 0.63	r = 0.77	NR
Garcia-Losa, 2001 [61]	IPSS	4 weeks	ICC = 0.76	NR	NR
Badia, 1998 [62]	IPSS	1 week	ICC = 0.87 (without item 8) r = 0.92	NR	Cα = 0.81
Cam, 2004 [52]	IPSS	NR	NR	No significant differences between raters.	NR
Hammad, 2010 [57]	IPSS	1-2 weeks	ICC = 0.88	NR	Cα = 0.85
Johnson, 2009 [55]	AUA SI	NR	NR	r = 0.75 (rho) k = p<0.001	NR
O'Connor, 2003 [60]	IPSS	1 week	r = 0.81	NR	NR

Okamura, 2009 [58]	IPSS	NR	NR	NR	Cα = 0.80 (women) Cα = 0.74, 0.79 (range between groups of men).
Quek, 2005 [59]	IPSS	1 week	ICC = >0.93	NR	Cα = 0.97
Basra, 2007 [74]	B-SAQ	4 weeks	k = 0.60 to 0.69	NR	Cα = 0.91

ICIQ-UI SF = International Consultation on Incontinence, Urinary Incontinence Short Form; IPSS = International Prostate Symptom Score; AUASI American Urological Association Symptom Index; B-SAQ – Bladder Control Self-Assessment Questionnaire; k = kappa; ICC = Intra-class Correlation Coefficient; r = correlation coefficient (Pearson product-moment correlation unless specified otherwise); Cα – Cronbach's Alpha coefficient; NR = not reported

table 5: Validity Results for Included Studies

Reference	Screening Questionnaire	Content Validity: Face/Logical validity	Criterion Validity: Concurrent validity	Criterion Validity: Predictive validity	Construct Validity: Discriminative validity	Construct Validity: Convergent/Divergent validity	Reference standard used
Lim, 2017 [64]	ICIQ-UI SF	Clinical experts and participants all agreed that all 3 versions of the ICIQ-UI SF (English, Chinese and Malay) were intelligible and covered all important domains.	NR	NR	Could discriminate between women with Stress UI and healthy continent controls ($p < 0.001$). ROC = 1.00 for all three versions (English, Chinese and Malay).	NR	NR
Avery, 2004 [69]	ICIQ-UI SF	Interviews with UK urology clinic attendees and reviews by clinical and social science experts indicated that items in the ICIQ were well interpreted and covered all important domains.	NR	NR	Could discriminate between men and women ($p < 0.001$) and between community and urology clinic patients ($p < 0.001$). *1	r = 0.29 to 0.86 (rho) (agreement between responses to ICIQ-UI SF and BFLUTS items). r = 0.24 to 0.58 (rho) (agreement between responses to ICIQ-UI SF and BFLUTS/ICSMaleSF items, assessing perceived causes of incontinence). r = 0.72 (agreement between ICIQ-UI SF QOL item and KHQ).	BFLUTS, ICSMaleSF and KHQ for assessing convergent validity.
Gotoh, 2009 [65]	ICIQ-UI SF	NR	Linear trends found between clinical severity measures (including 1-hr pad test and n of daily incontinence episodes) and ICIQ-SF scores.	NR	NR	r = 0.74 (rho) (agreement between ICIQ-UI SF scores and severity measure subscale of the KHQ). r = 0.68 (rho) (agreement between ICIQ-UI SF scores and; incontinence impact subscale of the KHQ; limitations in functional ability subscale of the KHQ; physical limitations subscale of the KHQ). r = 0.59 (rho) (agreement between ICIQ-UI SF scores and social limitations subscale of the KHQ). r = 0.55 (rho) (agreement between	1hr pad test and n of daily incontinence episodes for assessing concurrent validity. KHQ subscales for assessing convergent validity.

						ICIQ-UI SF scores and emotions subscale of the KHQ). *2	
Hashim, 2006 [67]	ICIQ-UI SF	NR	NR	NR	Could discriminate between the types of incontinence reported by males and females $\chi^2 = 35.4$, $p < 0.0001$ Could discriminate between individuals with mixed UI and those either with urge UI or stress UI (Multivariable regression) ($P < 0.0001$).	NR	NR
Rotar, 2009 [66]	ICIQ-UI SF	NR	k = 0.77 (agreement between responses to Q6 of ICIQ-SF and urodynamic test results).	NR	Could discriminate between types of UI reported by men and women. Significant differences found $p < 0.001$. *1	NR	Urodynamic testing for assessing concurrent validity.
Mary-Heck, 2018 [71]	ICIQ-UI SF	NR	Se, and Sp were 100% when compared to a reference measure which was a team of health experts who made diagnostic decisions .	NR	NR	NR	NR
Tubaro, 2006 [68]	ICIQ-UI SF	NR	Large variability between ICIQ-UI SF scores and reported results from a 72-hour voiding diary, although specific results were not provided.	NR	Could discriminate between cases and controls. Used a Wilcoxon two-sample test. $P < 0.001$.	r = 0.485 (rho) (agreement between ICIQ-UI SF item on QOL and item 6 of the SF-36).	SF-36 for assessing convergent validity.
Barry, 1992 [50]	IPSS	NR	NR	NR	Could discriminate between BPH patients and control patients. ROC = 0.82 to 0.94.	r = 0.85 (agreement between IPSS and Madsen-Iversen). r = 0.88 (agreement between IPSS and MMAP).	MMAP (Maine Medical Assessment Programme.), Boyarsky, and Madsen-Iversen, for assessing convergent

						r = 0.93 (agreement between IPSS and Boyarsky).	validity.
Badia, 1998 [62]	IPSS	NR	NR	NR	Could discriminate between BPH patients and control patients. ROC = 0.95 (whole instrument) ROC = 0.79 to 0.88 (individual IPSS Sp items).	r = -0.07 to 0.36 (rho) (agreement between IPSS Sp and EQ dimensions). r = -0.29 (rho) (agreement between IPSS Sp and EQ visual analogue scale). r = 0.14 to 0.72 (rho) (agreement between IPSS SP and PGWBI dimensions). r = 0.72 (rho) (agreement between IPSS Sp and item 8 of the IPSS).	EQ-5D, EQ-VAS, and PGWBI for assessing convergent validity.
Hammad, 2010 [57]	IPSS	NR	NR	NR	Could discriminate between BPH patients and control patients. ROC = 0.93 ROC = 0.79 to 0.90 (individual IPSS items).	r = 0.82 (rho) (agreement between IPSS Sp and item 8 of the IPSS)	NR
Okamura, 2009 [58]	IPSS	NR	NR	NR	NR	r > 0.33 (rho) (agreement between most voiding items and most storage items on IPSS). r ≤ 0.33 (rho) (agreement between different categories, i.e. voiding with storage items on IPSS).	NR
Basra, 2007 [74]	B-SAQ	The B-SAQ was developed by a European panel of experts in LUTD. Opinions from the expert panel and focus group interviews with patients concluded that the B-SAQ items were relevant.	NR	NR	NR	r = 0.46 to 0.54 (agreement between B-SAQ symptom scores and symptom severity scale of the KHQ) r = 0.79 (agreement between B-SAQ symptom scores and the incontinence impact domain of the KHQ). r = 0.81 (agreement between B-SAQ bother scores and the incontinence impact domain of the KHQ). k = 0.62 to 0.71 (agreement between individual items of the B-SAQ and individual items of the symptom	KHQ for assessing convergent validity.

						severity scale of the KHQ). k = 0.52 (agreement between overall B-SAQ scale and KHQ).	
Sahai, 2014 [73]	B-SAQ	NR	NR	NR	NR	r = 0.94 (agreement between B-SAQ symptom scores and B-SAQ bother scores). r = 0.86 (agreement between B-SAQ frequency symptom scores and individual symptoms in the symptom score domain of the KHQ) r = 0.85 (agreement between B-SAQ nocturia symptom scores and individual symptoms in the symptom score domain of the KHQ).	KHQ for assessing convergent validity.

ICIQ-UI SF = International Consultation on Incontinence, Urinary Incontinence Short Form; IPSS = International Prostate Symptom Score; B-SAQ – Bladder Control Self-Assessment Questionnaire; k = kappa; r = correlation coefficient (Pearson product-moment correlation unless specified otherwise; ROC = areas under their receiver operating characteristic curves; χ^2 = Chi square; NR = not reported; * 1 = Chi square test results not provided; *2 = Correlation for total scores for ICIQ-UI SF and KHQ not provided; KHQ = Kings Health Questionnaire; BFLUTS = British Female Lower Urinary Tract Symptoms questionnaire; ICSMaleSF = The International Continence Society Male Questionnaire Short Form; SF-36 = 36-Item Short Form Survey Instrument; MMAP = Maine Medical Assessment Programme; EQ-5D = EuroQol-5D; EQ-VAS = EuroQol visual analogue scale; PGWBI = Psychological General Well-Being Index; Se = sensitivity; Sp = specificity.

Table 6: Critical Appraisal of Included Studies

Question	1(r, v)	2 (v)	3 (r)	4 (r)	5 (r)	6 (v)	7 (r)	8 (v)	9 (r, v)	10 (v)	11 (r, v)	12 (r, v)	13 (r, v)	14 (r, v)	15 (r, v)	% yes
Studies measuring reliability																
Lim (2017) [64]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	Y	78
Avery (2004) [69]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	Y	Y	N	Y	N	67
Barry 1992 (a) [50]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	Y	78
Barry 1992 (b) [51]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	Y	78
Barry 1995 [56]	Y	NA	NA	NA	Y	NA	Y	NA	Y	NA	Y	Y	Y	Y	Y	90
Basra 2006 [74]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	Y	Y	Y	Y	N	78
El Din 1999 [63]	Y	NA	N	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	N	60
Johnson, 2009 [55]	Y	NA	Y	NA	N	NA	Y	NA	Y	NA	Y	Y	Y	Y	Y	90
O'Connor 2003 [60]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	Y	Y	N	Y	N	67
Cam, 2004 [52]	Y	NA	N	NA	Y	NA	Y	NA	Y	NA	Y	Y	Y	Y	N	80
Okamura,	Y	NA	NA	NA	N	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	90

2009 [58]																
Garcia-Losa, 2001 [61]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	Y	Y	Y	Y	Y	90
Rotar, 2009 [66]	Y	NA	Y	NA	N	NA	Y	NA	Y	NA	Y	Y	Y	Y	Y	90
Badia, 1998 [62]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	Y	78
Gotoh, 2009 [65]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	N	67
Hammad 2010 [57]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	Y	Y	Y	Y	N	78
Quek 2005 [59]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	N	Y	Y	Y	Y	78
Tubaro 2006 [68]	Y	NA	NA	NA	N	NA	Y	NA	Y	NA	Y	Y	Y	Y	N	78
Hashim 2006 [67]	Y	NA	NA	NA	N	NA	Y	NA	N	NA	N	Y	Y	Y	N	56
Kurzawa 2018 [70]	Y	NA	NA	NA	NA	NA	NA	NA	Y	NA	N	Y	Y	Y	Y	86
Uren, 2017 [82]	N	NA	NA	Y	Y	NA	N	NA	Y	NA	Y	Y	Y	Y	Y	80
Studies measuring validity																
Lim (2017) [64]	Y	NA	NA	NA	NA	NA	NA	NA	Y	NA	N	Y	Y	Y	Y	86
Avery (2004) [69]	Y	N	NA	NA	NA	Y	NA	Y	Y	N	Y	Y	N	Y	N	64
Barry 1992 (a) [50]	Y	Y	NA	NA	NA	Y	NA	Y	Y	Y	N	Y	Y	Y	Y	91
Barry 1992 (b) [51]	Y	Y	NA	NA	NA	Y	NA	Y	Y	N	N	Y	Y	Y	Y	82

Basra 2006 [74]	Y	Y	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	Y	Y	N	91
Sahai 2014 [73]	Y	Y	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	Y	Y	Y	100
Okamura, 2009 [58]	Y	NA	NA	NA	NA	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	100
Rotar, 2009 [66]	Y	Y	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	Y	Y	Y	100
Badia, 1998 [62]	NA	Y	NA	NA	NA	Y	NA	Y	Y	Y	N	Y	Y	Y	Y	90
Gotoh, 2009 [65]	Y	Y	NA	NA	NA	Y	NA	Y	Y	Y	N	Y	Y	Y	N	82
Hammad 2010 [57]	Y	N	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	Y	Y	N	82
Tubaro 2006 [68]	Y	Y	NA	NA	NA	Y	NA	Y	Y	Y	Y	Y	Y	Y	N	91
Hashim 2006 675]	Y	NA	NA	NA	NA	Y	NA	NA	N	NA	N	Y	Y	Y	N	63
Mary-Heck, 2018 [71]	Y	Y	Y	NA	NA	Y	NA	Y	Y	Y	N	Y	Y	Y	N	83

Question1= Did the authors give a detailed description of the sample of subjects used to perform the (index) test?; Question 2 = Was the reference standard explained?; Question 3 = If inter-rater reliability was tested, were raters blinded to the findings of other raters?; Question 4 = If intra-rater reliability was tested, were raters blinded to their own prior findings of the test under evaluation?; Question 5 = Was the order of the questions/tests varied?; Question 6 = Was the time period between the reference standard and the index test short enough to be reasonably sure that the target condition did not change between the two tests?; Question 7 = Was the stability (or theoretical stability) of the variable being measured taken into account when determining the suitability of the time interval between repeated measures?; Question 8 = Was the reference standard independent of the index test?; Question 9 = Was the execution of the (index) test described in sufficient detail to permit replication of the test?; Question 10 = Was the execution of the reference standard described in sufficient detail to permit its replication?; Question 11 = Were withdrawals from the study explained?; Question 12 = Were the statistical methods appropriate for the purpose of the study? Question 13 = Were subjects selected either randomly or consecutively? Question 14 = Was the number of subjects either >50 or was a sample size calculation provided?; Question 15 = Did subjects give consent prior to testing/participating?; r = reliability; v = validity; NA = not applicable.